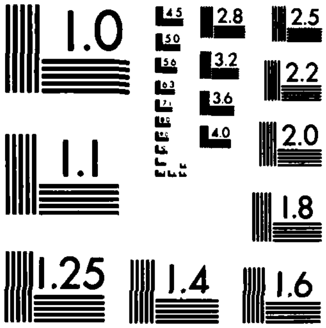
AD-A132 200 EVALUATOR BIAS IN THE MARINE CORPS COMBAT READINESS EVALUATION SYSTEM (MCCRES) ITS IDENTIFICATION AND CONTROL(U) NAVAL POSTGRADUATE SCHOOL MONTEREY CA 1/1

UNCLASSIFIED G M WHEELER JUN 83 F/G 5/9 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

2

NAVAL POSTGRADUATE SCHOOL

Monterey, California

THESIS

EVALUATOR BIAS IN THE MARINE CORPS COMBAT
READINESS EVALUATION SYTEM (MCCRES)
ITS IDENTIFICATION AND CONTROL

by

George M. Wheeler
June, 1983

Kenneth Euske
Thesis Co-Advisor Joseph Mullane

Approved for public release; distribution unlimited

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		**READ INSTRUCTIONS BEFORE COMPLETING FORM**
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Evaluation Bias in the Marine Corps Combat Readiness Evaluation System (MCCRES) Its Identification and Control		5. TYPE OF REPORT & PERIOD COVERED Master's Thesis June, 1983
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) George M. Wheeler		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940		12. REPORT DATE June, 1983
		13. NUMBER OF PAGES 64
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Combat Readiness Evaluation, MCCRES,
Bias,

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Marine Corps Combat Readiness Evaluation System (MCCRES) was designed to provide timely and accurate information concerning the ability of active and reserve forces to carry out assigned combat missions. To provide this information, units are subjected to simulated combat problems and their performance is observed by expert evaluators from within the Marine Corps. Though these evaluators are considered experts in their fields, they may inject bias into their evaluations causing an inaccurate (CONT)

Abstract (Continued) Block # 20

combat readiness rating for the unit observed.

Analysis of the MCCRES reveals three main areas where evaluator bias may appear: senior evaluator influence, other evaluator bias and interpretation of the mission performance standards used to conduct the evaluation. To alleviate these problems, three actions are explored: evaluator training, evaluator testing and quantification of the mission performance standards.

Approved for public release; distribution unlimited.

Evaluator Bias in the Marine Corps Combat Readiness Evaluation System (MCCRES) Its Identification and Control

by

George M. Wheeler
Captain, United States Marine Corps
B.S.A.E., United States Naval Academy, 1976

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN INFORMATION SYSTEMS

from the

NAVAL POSTGRADUATE SCHOOL
June 1983

Author: George M. Wheeler

Approved by: K J Euske
Thesis Co-Advisor

Mullen Jr
Thesis Co-Advisor

Richard L. Elster
Chairman, Department of Administrative Sciences

Kneale T. Marshall
Dean of Information and Policy Sciences

ABSTRACT

The Marine Corps Combat Readiness Evaluation System (MCCRES) was designed to provide timely and accurate information concerning the ability of active and reserve forces to carry out assigned combat missions. To provide this information, units are subjected to simulated combat problems and their performance is observed by expert evaluators from within the Marine Corps. Though these evaluators are considered experts in their fields, they may inject bias into their evaluations causing an inaccurate combat readiness rating for the unit observed.

Analysis of the MCCRES reveals three main areas where evaluator bias may appear: senior evaluator influence, other evaluator bias and interpretation of the mission performance standards used to conduct the evaluation. To alleviate these problems, three actions are explored: evaluator training, evaluator testing and quantification of the mission performance standards.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

A. PURPOSE

The purpose of this paper is to examine the Marine Corps Combat Readiness Evaluation System (MCCRES) to discover if the system is susceptible to biases which may cause the results of evaluations to inaccurately reflect the combat readiness of evaluated units. To guide research, two specific questions are posed:

1. Can factors of the MCCRES evaluation which are subject to evaluator bias be identified?
2. How can these factors be controlled or controlled for?

B. BACKGROUND

The Marine Corps Combat Readiness Evaluation System was designed to provide timely and accurate information concerning the ability of operating units of the Marine Corps, both active and reserve, to carry out assigned combat missions. The system uses "expert" evaluators from various specialty areas to observe and grade simulated combat operations. Aggregating these evaluations provides an overall view of a unit's readiness for combat, and feedback from the evaluation allows the unit commander to identify and correct potentially problematic areas within his command.

Though the MCCRES is relied upon as a standard against which units are judged, the readiness grade received could be more dependent upon the evaluator than the actual task performance being graded. By controlling or controlling for evaluator bias, a more uniform standard by which to judge combat readiness can be realized.

C. SCOPE AND METHODOLOGY

This thesis views the MCCRES as an information system and explores areas where evaluator bias (input) can cause ratings (output) to reflect the evaluator's opinion rather than the mission performance of the evaluated unit. Two major topics are researched:

1. Evaluation--Its major approaches and principles
2. Evaluators--Their sources and typical errors

These areas are related to the MCCRES and methods of controlling or controlling for evaluator bias are developed.

The research consists of a detailed literature search in the area of evaluation science. Methods for the reduction or control of evaluator bias are explored for use in the context of the MCCRES.

II. EVALUATION

This chapter addresses the evaluation process, presenting definitions, purposes and principles of evaluation, and explores some currently used approaches for conducting evaluations. The questions of what to evaluate and when to evaluate are also investigated.

The terms goal and objective are used throughout this and succeeding chapters. Objectives refer to long range statements of purpose within the organization. They generally can not be specifically stated and need not be attainable in the immediate future. Alternatively, goals are more readily attainable in the short run and are specifically stated. They can appear as written statements which guide an organization's operations, and are a standard against which performance can be measured.

A. DEFINITION AND PURPOSE OF EVALUATION

1. Definition of Evaluation

There are many definitions of the term evaluation. Rather than select a single author's definition, two observations and two definitions of evaluation are presented here to show both the similarities and differences encountered in the field of evaluation research. These definitions and observations are given in order from simple to rigorous.

The first, more an observation than a definition, is from E.R. House:

> At its simplest, evaluation leads to a settled opinion that something is the case. It does not necessarily lead to a decision to act in a certain way, though today it is often intended for that purpose....Evaluation leads to a judgement about the worth of something. [Ref. 1:p.18]

The second observation about evaluation, in particular the evaluation of a process, is that its scope "is confined to assessing what a particular program has accomplished in meeting its immediate objectives...," and assessing the "workability " of a program [Ref. 2 :p.11].

Henry W. Rieken's definition looks upon evaluation as " the measurement of desireable and undesireable consequences of an action that has been taken in order to forward some goal that we value." [Ref. 3 :p.54]

Finally, the definition presented by Stufflebeam et al., is that "...evaluation is the process of delineating, obtaining, and providing useful information for judging decision alternatives." [Ref. 4 :p.40]

There are two factors common to each of the preceeding observations and definitions. First, evaluation is concerned with making a judgement or assessment about something. Second, that judgement can be made in terms of some goal or objective. These two factors are used as a basis for a definition of evaluation developed in the next section.

2. Purpose of Evaluation

Using the above descriptions of evaluation, the purpose of evaluation can be examined. Stufflebeam et al., stated simply that "The purpose of evaluation is not to prove but to improve." [Ref. 4] Combining this statement with the ideas set forth in defining evaluation, we may look at evaluation as a judgement of something, say a program, with the purpose of improving the current attainment of that program's goals or objectives. This position, though, seems to make evaluation a method of program improvement rather than a tool to help achieve this end. The judgement made may indicate some action which should be taken to improve the organization's goal attainment, but the judgement in and of itself does not cause the organization's goal attainment to improve. As such, the evaluation is a tool for program improvement. Evaluation as a tool for decision making is brought out by Anderson and Ball. Their use of the phrase "...to contribute to decisions..." [Ref. 5] in describing evaluation makes clearer the idea that evaluation is a tool rather than an end in itself.

If the above purposes of evaluation are accepted, then we may wish to form a new definition of evaluation. This definition takes into account evaluation's purpose. Aggregating the previously cited authors' opinions and definitions we may look at evaluation as a judgement of some program with the purpose of contributing to decisions concerning the current attainment of that program's goals or objectives.

B. PRINCIPLES OF EVALUATION

There appears to be a general acknowledgement among authors of evaluation literature that a group of principles exists which governs the conduct of evaluations. Tracey [Ref. 6] listed six principles which may be found in various forms in the writings of other authors [Ref. 1, 4, 5, 8, 9]. Evaluation must:

1. *Be conducted in terms of purposes, that is the objectives must be known.* If the objectives are not known, the evaluation effort cannot measure how well they are being attained.
2. *Be cooperative.* Cooperation of all organizational levels is essential. Without free communication, evaluation results will not reach all parties, diluting their usefulness.
3. *Be continuous.* Evaluation must be an on-going process to accurately track performance and aid planning in light of current objective attainment.

4. Be specific. Generalizations are not as useful as specific information in providing performance information.

5. Provide means and focus to appraise self, practice and product. The evaluation must provide information of sufficient quantity and specificity to evaluate not only the program output, but the mechanism of converting inputs to output and the individuals' performance within the mechanism.

6. Be based on uniform and objective methods and standards. Methods and standards which change from one evaluation to the next destroy trust and leave those being evaluated questioning how they should perform their work tasks. [Ref. 6:p.14-15]

C. APPROACHES TO EVALUATION

How does one approach or categorize evaluation? The following section discusses eight approaches to or categories of evaluation forwarded by House [Ref. 1:p.21-43].

1. The Systems Analysis Approach

The systems analysis approach defines a small number of output measures and attempts to relate differences in programs to variations observed in the variables. The data acquired through this observation is quantitative in nature. Correlational analysis or other statistical methods are used to relate the output measures to the programs being evaluated. This method is widely used in the Department of Health, Education and Welfare in evaluating federal social welfare programs.

An example is the Office of Economic Opportunity (OEO) evaluation of the Neighborhood Health Center (NHC) program. The OEO defined five areas of interest to be investigated in determining the impact of the NHC's. These areas of interest were:

1. Success of the NHC's in providing comprehensive health care to the poor.
2. Patient reaction to the care received at the NHC's.
3. Degree of implementation of comprehensive and continuous family care at the NHC's.
4. Functional and organizational comparison of the NHC's.
5. Antipoverty consequences of NHC services. [Ref. 7:p.107-121]

The NHC program was evaluated according to the attainment of the objectives which relate to the five specified interest areas.

One problem which may be seen with this approach is ensuring the output measures selected truly reflect the organization's goals. If the selected measures do not accurately reflect those goals, the outcome of this approach may be of limited use.

2. The Behavioral-Objectives (Or Goal-Based) Approach

This approach, popularized in business and government organizations as management by objectives, uses the stated goals of a program as the output measure and evaluates program success by the attainment of these goals. It can be seen that this method of evaluation addresses only the issue of program effectiveness, providing no information on program efficiency. In this sense, effectiveness is a measure of the extent to which an organization's objectives are achieved. Efficiency refers to the cost of converting program inputs to outputs, that is, the cost of objective achievement. An early advocate of this behavioral-objective approach was Tyler [Ref. 8] who advanced this method for evaluating educational goals in terms of student behaviors.

Peter F. Drucker popularized the term "management by objectives" in his book The Practice of Management [Ref. 9]. Implementation of management by objectives (MBO) forces individuals and organizations to define specific areas of responsibility in terms of measureable expected results, called objectives. Performance is determined by comparing objective attainment against the objectives stated. The popularity of the approach can be seen in its widespread use. A 1976 study showed 41 percent of the hospitals surveyed used MBO and another 33 percent were planning to start in the near future [Ref. 10:p.8-11]. MBO is used not only as an evaluation approach, but as a means of planning, coordination, communication and control. An advantage is the explicit statement of objectives which let workers know their specific duties and encourages communication between workers and supervisors relating to job performance. A major disadvantage is the problem of specifying behaviors rather than performance. Specific objectives are very measureable, but behaviors are not necessarily measureable in the context of contributing to goal attainment. Waks [Ref. 1:p.487] argued that "...acting with purpose..." is not equivalent to "...taking means to a well defined end." In other words, though a specified behavior may be observed, it does not follow that this behavior leads to a desired objective.

3. The Decision-Making Approach

As an earlier definition of evaluation implied, evaluation is closely related to decision-making. The decision-making approach holds that an evaluation is structured according to the decisions which must be made. It assumes that the decision-maker's concerns are the significant areas the evaluation must address. By structuring the evaluation in this manner, the results should be of greater use to the decision-maker. This approach relies heavily on survey methods such as interviews and questionnaires.

Stufflebeam et al. [Ref. 4], whose previously cited definition of evaluation includes the idea that evaluation is to provide information for judging decision alternatives, is an advocate of this approach in the field of education. The evaluation is structured with respect to the decision-makers' concerns and position in the organization, and specific evaluation subtasks are identified and assigned. The results of these subtasks are aggregated and communicated to the decision-maker in order to aid in the decision process. [Ref. 4] This approach relieves the evaluator from having to guess the audience of the evaluation, thereby providing structure for the entire evaluation effort. On the other hand, this approach assumes that the decision maker's goals are the same as those of the entire organization, which may or may not be the case.

4. The Goal-Free Approach

Each of the previously discussed approaches involved program evaluation in terms of program goals and specific goals for the evaluation. The goal-free approach seeks to conduct evaluation in terms of program goals without reference to the goals for the evaluation, indeed, the evaluator is purposely kept unaware of these goals so as not to be biased by them.

Scriven [Ref. 11], a leading proponent of this school of thought, feels that the goal-free approach is a valid method of reducing bias in evaluation, since knowledge of evaluation goals can influence the evaluator. For example, an evaluator who is tasked with conducting a performance evaluation of an employee with the explicit intent of determining whether the employee should be terminated may deliver a different evaluation if the intent is not stated. In the former instance, evaluator knowledge that his evaluation may result in a worker losing his job may bias the outcome of the evaluation. By being unaware of the evaluation intent, the latter situation may result in a more accurate representation of the worker's performance.

This approach is widely used in the area of consumer product evaluations. Various consumer organizations regularly evaluate products placed in the market without knowledge of the manufacturers goals. These evaluations stress standards and criteria which they (the consumer organization) feel are beneficial to the consumer. One main problem to overcome in this approach is the choice of evaluators. Scriven [Ref. 11] sees evaluators as experts, able to eliminate and prevent both self-bias and bias of others from impacting on the evaluation. A variety of techniques, such as codes of ethics or double-blind experiments, are available to assist the evaluator in eliminating bias.

5. The Art Criticism Approach

This approach relies upon the critic to make judgement on a program much the same way an art critic would judge a fine painting. Though opinions on specific details may vary, there is generally a consensus among critics of a certain endeavor as to what constitutes a notable work. This implies an extensive base of common knowledge among those eligible to conduct such criticism.

Eisner makes a distinction between connoisseurship and criticism. While connoisseurship is "recognizing and appreciating the qualities of the particular" it requires no public disclosure or judgement. Criticism necessarily encompasses connoisseurship. "Criticism is the art of disclosing the qualities of events or objects that connoisseurship perceives." [Ref. 12 :p.197]

The key purpose of criticism is to increase awareness of a subject area and convey judgements in terms of criteria which are accepted among those knowledgeable in that area. It allows the uninitiated to gain an appreciation for that area through the critic's knowledge. Though generally associated with art, literature and other basically creative areas, the art criticism approach to evaluation has been applied to the field of education with some success.

A key problem with this approach is generating acceptance of the critic's criteria for judging a program. A critic may possess extensive knowledge in his field, but if the audience of his evaluaton is not receptive, his criticism is not likely to carry much weight.

6. The Professional Review (Accreditation) Approach

The professional review approach has some distinct parallels with the art-criticism aproach immediately above. Professional review relies upon expert opinion concerning generally accepted standards of performance in evaluating a particular area. The standards here, though, are usually more easily quantified, leading to a more structured approach in the evaluation. Professional review also is apt to use many members, organized as an accreditation or review board to conduct the evaluation. Standards and measurement criteria are determined by the professionals themselves as they are accepted as the experts in their fields. This approach produces an evaluation of professionals by professionals and its outcomes are not easily influenced by the layman.

7. The Quasi-Legal (Adversary) Approach

One of the long standing approaches for evaluating and policy-making is the quasi-legal approach. It is an approach to evaluation which closely imitates legal procedures. Information, or 'evidence', concerning a program is obtained from 'witnesses', much as testimony is received in a court of law. Information both for and against a particular program is presented, and great care is exercised to ensure that all pertinent information is received after which a panel of evaluators weighs the evidence heard and can reach a decision as to the worth of the program. Examples of this approach abound in today's government, ranging from local school board decisions on grade school curricula through presidentially appointed panels like the Warren Commission which investigated the assassination of President Kennedy.

This approach does not rely only on expert evaluators as have several previous approaches. Additionally it not only accepts but encourages personal bias and opinion in those providing information. As Wolf notes:

> The ultimate evidence which guides deliberation and judgement includes not only the 'facts', but a wide variety of perceptions, opinions, biases, and speculations, all within a context of values and beliefs. [Ref. 13:p.21]

The ultimate goal of this approach is to reach a definite conclusion on some issue. Its conclusions will address absolutes, such as 'Is the program meeting its goals' rather than matters of degree, as 'To what extent are our goals met'.

8. The Case Study (or Transaction) Approach

This approach is widely used and accepted in organizational studies. It focuses on program processes and interactions, both within and outside the program, with the intent of giving the reader of the case study a greater appreciation of the program's workings. This approach commonly presents interviews with people in the program and observations made by the interviewer at the program site in the form of a case. The case can be examined by evaluators and conclusions reached through discussions and sharing of ideas among the evaluators. The case study and its conclusions are aimed at the reader who does not possess a great knowledge of the evaluation area as a means of increasing his/her understanding by illustrating how others view the program being evaluated. This approach allows the reader to more fully understand the internal workings of the program and how program inputs are converted to outputs.

A major problem with this approach can be ensuring confidentiality for the members upon which the case study was based. Case study authors may have difficulty disguising all of the personalities involved in a case. Another problem which may be encountered is representing fairly the great diversity of actions and opinions which a large case study may entail. A complicated case with many personal interactions can require a tremendous editorial effort to ensure that it is accurate and understandable.

9. Summary

The above approaches are certainly not all inclusive, nor can all approaches to evaluation be expected to fit into these eight categories. They are intended to show the variety of approaches available in conducting evaluations. Though the overall purpose of evaluation may be the same, that is providing information to aid in decision making, different situations may call for different approaches to provide necessary information. The eight approaches show that techniques can be chosen to fit evaluation to evaluator skill (quasi-legal vs. professional review approaches), program objectives (system analysis vs. behavioral-objectives approaches), or even to ignore evaluation objectives (goal-free approach).

D. WHEN TO EVALUATE

Stufflebeam et al. [Ref. 4] provide a view of evaluation which investigates when in the program life cycle evaluation is to take place. They have defined four types of evaluation--context, input, process, and product evaluation--which serve functions from program inception through final impact on the system in which the program operates. Each evaluation type is explained briefly below.

1. Context Evaluation

Context evaluation is used in the planning process with the intent of identifying unmet goals or unused opportunities and identifying problems which prevent the goals from being met or the opportunities from being used. This problem identification leads to formulation of program objectives which are used as yardsticks against which program performance is measured. Stufflebeam et al. [Ref. 4] further identify two modes of context evaluation: contingency and congruence. The contingency mode looks outside the system for factors which may yield improvements within. Typically, if-then type questions relating outside factors to objectives are asked--if our manning level is reduced by 20% then can we carry out our mission? If research costs continue to rise, then is our present budget adequate? Congruence mode is a comparison between goals and actual performance. This mode informs the organization as to its goal attainment. As opposed to contingency mode, congruence mode looks only within the system in question to provide evaluation data.

2. Input Evaluation

Input evaluation is concerned with the use of available resources in obtaining objectives formulated in context evaluation. It is useful in providing information to be used in structuring the program, and its output can be compared to a cost/benefit analysis with resource usage as the cost and goal attainment as the benefit. Besides program structuring, input evaluation also helps address such problems as the need for additional resources and other general strategic decisions.

3. Process Evaluation

Process evaluation begins after program approval and implementation. Process evaluation analyzes the program process as it is operating to provide information on whether the process is working as designed. Stufflebeam et al. [Ref. 4] point out that this type of evaluation is particularly important early in program implementation, when firm output information is not yet available. It allows the organization to measure how well it is carrying out the program plan.

4. Product Evaluation

Product evaluation provides information on goal attainment, how well the stated objectives are met. It is a major input to decisions which would modify the program after implementation.

The view provided by Stufflebeam et al. [Ref. 4] should not be regarded as an evaluation approach different from those listed by House [Ref. 1], but as an expansion of those approaches. Each of the eight approaches could be structured to look specifically at input, context, process or output though, as implied earlier, the different approaches may not be equally effective in providing information in these four areas. The Stufflebeam et al. view can be seen as helping determine the timing of evaluations, using one of House's approaches, to provide information on specific portions of a program's life-cycle.

E. SUMMARY

This chapter has focused on the many ideas and approaches available in evaluation science. Definitions of evaluation and its purposes were presented to show the similarities and differences that exist among authors of evaluation literature and a definition of evaluation was formed. The definition looked upon evaluation as a judgement of some program with the purpose of contributing to decisions concerning the current attainment of that program's goals or objectives. Six principles for evaluation were also presented, demonstrating how and when evaluation should be conducted and what kind of information should be provided by the evaluation.

The basic concepts of evaluation were expanded by investigating eight approaches which are available to evaluators. These approaches provide different evaluation structure depending on the type of information desired from the evaluation or the different evaluation assets available. Finally, a view of evaluation which addresses when to perform evaluation was added to the eight evaluation approaches.

With this grounding in the fundamental ideas of evaluation, the next chapter will focus on the evaluator's roles and responsibilities, and some problems associated with evaluation. The evaluator's implementation of the above principles and methods can greatly influence the eventual outcome of the evaluation.

III. EVALUATORS

> The ideal rater who observes and evaluates what is important and reports his judgement without bias or appreciable error does not exist, or if he does, we don't know how to separate him from his less effective colleagues. [Ref. 14:p.7]

Though the above statement may be true, many steps have been taken in evaluation science to identify competent evaluators and improve performance of evaluators in general. This chapter looks at the evaluator, beginning with a discussion of objectivity and validity as they relate to evaluation. Who performs evaluations and whether they come from within or outside the organization is investigated, with advantages and disadvantages presented for each evaluation source. A discussion of the kinds of errors evaluators typically make is presented along with sources which may cause these errors. The chapter closes with a discussion of several methods for reducing the amount of errors evaluators may bring into their evaluations, ranging from training the evaluator to improving the tools the evaluator uses in performing evaluation.

A. OBJECTIVITY

Objectivity, in the context of evaluation, is the ability to observe something only as it physically exists without the inclusion of personal feelings about the object. For example, the statement 'Joe is six feet tall' would be considered more objective than saying 'Joe is a giant'. The former could be adequately demonstrated using a tape measure, while the latter is largely dependent upon the particular observer's concept of what is giant and what is not. As House points out:

> Objectivity is often equated with agreement among observers. Agreement is accomplished by having externalized, specified procedures for observation. By this definition objectivity is achieved by having observers agree on what they see--replication of observation. [Ref. 1:p.215]

House calls this the quantitative notion of objectivity. The concept of reliability in observation closely parallels this quantitative notion. Reliability is based on the ability to replicate observations. That is, if a particular observation of an object can be replicated, that observation is assumed to be reliable.

B. VALIDITY

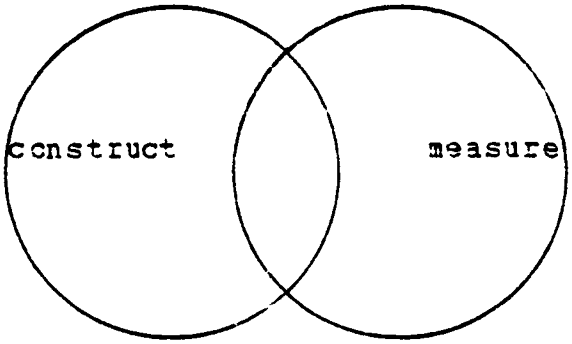
The concept of validity is important to evaluation. If an observation does not accurately reflect the qualities of an object one wishes to measure, a 'true' evaluation of that object may be impossible. Scriven [Ref. 15] addresses the concept of validity by bringing out a feature which he calls the qualitative sense of objectivity. He argues that, taken in the extreme, the quantitative notion of objectivity confuses the method of verification with 'truth'. An observation may be widely agreed upon and replicateable, but how closely does it represent reality? How 'good' is the observation? To illustrate, Scriven cited the incident of a television receiver evaluator observing picture quality. The evaluator used a mechanical device to measure decibel gain of the receivers, though there was little correlation between decibel gain and picture quality. The observations obtained were able to be replicated and the results widely agreed upon but they did not really relate to picture quality. In this case, the evaluation was quantitatively reliable but lacked quality. [Ref. 15] The issue of evaluation quality is commonly referred to as validity.

As a method of relating observations to objects we wish to evaluate, Cummings and Schwab [Ref. 16] suggest the concept of construct validity. A construct is a mental image we have of something, the way we perceive something. Validity, in this context, refers to the correlation between our mental image and some measure of it. In the previous example, there was little correlation between decibel gain of the television receivers and quality of the picture hence there was little construct validity. A different measure which more closely corresponds to our mental image of picture quality could be chosen. The closer the measure chosen corresponds with our mental image of something, the greater the construct validity. A different measure such as viewer satisfaction will have varying degrees of construct validity according to how closely it compares with our mental image of picture quality.

To better illustrate the concept of construct validity, consider Figure 3.1. As shown, the left circle represents some construct we are interested in and the right circle represents some measure of that construct. Ideally, there would be complete overlap of the circles representing a total correlation between the construct and the measure used. There are two general reasons that the two circles do not completely overlap--measurement deficiency and measurement contamination [Ref. 16].

Measurement deficiency occurs when the measure fails to take into account all of the factors present in our construct. For example, a measure of a data processing department's performance which accounted for quantity of output but neglected quality and timeliness would probably be considered deficient.

Measurement contamination, in contrast to measurement deficiency, occurs when the measure takes into account factors which fall outside our construct. If our measure of



Adapted from [ref.17 :p.75]

Figure 3.1 Deficiency and Contamination.

the data processing department's performance includes items such as corporate sales or top management's perceptions of the department, the measure is likely to be contaminated.

It may be seen that both deficiency and contamination in measurement of constructs adversely affect construct validity. If our measures do not contain all the factors pertinent to our construct, or if the measures contain factors outside our construct, it is unlikely that the measures will accurately reflect the mental image of the construct. Both of these circumstances, then, decrease construct validity.

C. ERRORS

There are a number of errors which evaluators may commit during the evaluation process. Cummings and Schwab [Ref. 16] discuss these errors in two main groups- variable error and constant error. These two groups are explained below, with examples.

1. Variable Error

Variable error is evaluator disagreement which manifests itself as differences in the scores of specific items of an evaluation. It may take two forms--disagreements between evaluators and disagreements over time.

a. Disagreements between evaluators

Suppose two evaluators, A and B, have observed five workers performing their jobs and rated the workers' performance on a scale of 0 (poor performance) to 10 (high performance). The ratings are shown in Table I. Note that there is total rating agreement only on worker 4 and the other ratings differ from 1 to 4 units.

TABLE I

Evaluator Ratings

	RATINGS	
WORKERS	EVALUATOR A	EVALUATOR B
1	5	3
2	7	8
3	3	7
4	9	9
5	4	0

Taking the ratings obtained from A and B, we now wish to plot the scores, with evaluator A's rating representing the X-component of our plot and evaluator B's ratings representing the Y-component of the plot. The result is a graph as shown in Figure 3.2. The straight line extending from the origin and rising from left to right represents total agreement between the evaluators. The distance of each worker's score from the total agreement line is a measure of the disagreement between the evaluators. A linear correlation coefficient may be calculated which expresses the amount of agreement between the evaluators. Values for the linear correlaton coefficient may vary from -1.0 (highly negative correlation, meaning that high values for the X-component tend to go with low values for the Y-component and low values for the X-component tend to go with high values for the Y-component) to +1.0 (highly positive correlation, meaning that high values for the X-component tend to go with high values for the Y-component and low values for the X-component tend to go with low values for the Y-component), with a value of 0.0 indicating no correlation (no predictable pattern). In this example, the linear correlation coefficient is 0.6 indicating some positive correlation between evaluators A and B. A value in the range of 0.8 to 0.9 would tend to indicate a strong correlation between A and B. High correlation does not, however, guarantee a valid rating. It simply shows that A and B agree on what they have observed. Both A and B may be wrong in their ratings of worker 4, but their agreement would provide some confidence that their rating was correct.

Two methods which can reduce disagreement between evaluators are reduction or elimination of subjectivity in measurement instruments and ensuring evaluator familiarity with the job being evaluated. The former method reduces disagreements by relieving the evaluator of interpreting subjective measures. By using more objective evaluation measures, evaluator bias is less likely to be accidentally introduced [Ref. 20 :p.46]. Ensuring evaluator familiarity with the job being evaluated increases the likelihood of evaluating job factors which correlate highly with job performance.

b. Disagreements Over Time

Disagreements over time pertain to disagreements in evaluations made by one evaluator at different points in time. Suppose that, in the example of disagreements between

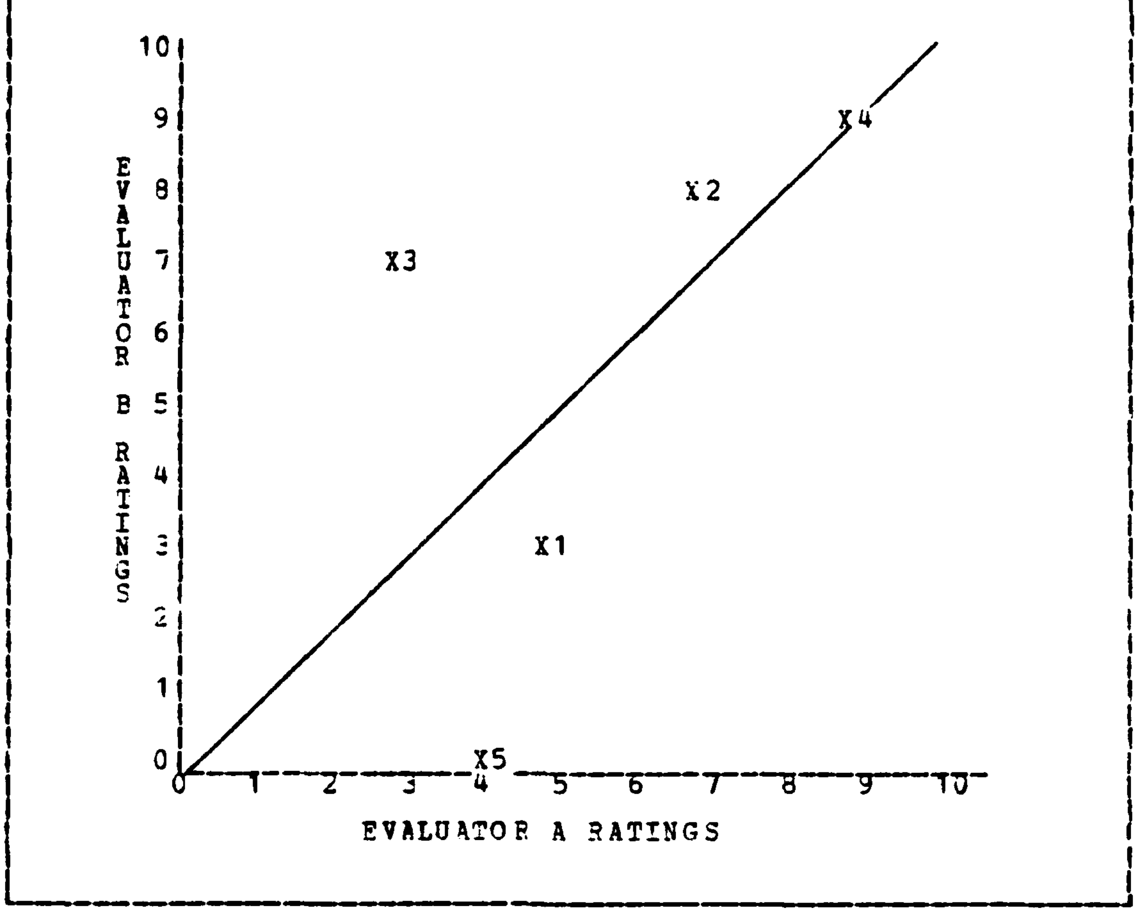


Figure 3.2 Evaluator Disagreements.

evaluators, evaluator A's ratings represented an evaluation performed by A at time 1 and that evaluator B's ratings represented an evaluation performed by A at time 2. Calculation of the linear correlation coefficient would then measure how well evaluator A's ratings agree over time.

Using disagreements over time as a measure of construct validity is generally not as desireable as using disagreements between evaluators. The reason for this is that differences in evaluations made at different points in time may be due to performance improvement or degradation of those being evaluated. The low correlation coefficient obtained from a comparison of evaluations made on a worker whose performance has changed markedly over time may be mistakenly taken to mean the construct is not valid. For this reason, correlation coefficients obtained by comparing two or more evaluators' ratings are a better measure of construct validity [Ref. 16]. A method of reducing disagreements over time, discussed later, is testing potential evaluators and choosing those who demonstrate little of this error.

2. Constant Errors

Where variable errors tend to create differences between evaluations, constant errors tend to cause spurious similarities. Constant error takes three forms--halo error, central tendency and leniency.

a. Halo error

Halo error occurs when an evaluator fails to differentiate among individual items or dimensions in his evaluation, but evaluates on the basis of his overall impression. The boss who observes only an employee's written work but rates the employee high in areas such as initiative and personal relations has made a halo error.

b. Central tendency

Central tendency is the tendency for evaluators to rate all dimensions of an object near the middle of the evaluation scale, avoiding the extremes.

c. Leniency

This error is committed when an evaluator tends to rate all objects too high. The 'easy grader' consistently delivers inflated rating marks. The opposite error, that of rating all objects too low is called strictness.

Evaluator training in the area of constant error is a useful technique in reducing these errors. A discussion of this technique is presented in a later section.

D. EVALUATION SOURCES

Evaluators may come from many places within and outside an organization. Though evaluations by superiors are very common, alternative sources of evaluation exist--peer, subordinate, self and disinterested party or outside evaluators.

1. Superior Evaluators

Evaluations by superiors are a widely used method in today's organizations. Superiors are chosen for many reasons, such as job experience, familiarity with subordinate positions and job skills, even tradition. Superiors are often the logical choice as evaluators, for their position in the organizational hierarchy is such that they determine to a great extent the incentive and reward system for their subordinates. As such, their evaluations of subordinates may lead to direct reward or punishment without passing through another level of hierarchy and this immediate evaluation-incentive tie keeps subordinates appraised of their performance.

Some problems can exist with supervisor evaluations. First, if the subordinate being rated does not work directly for the evaluating superior or if there is substantial physical separation of the supervisor from the subordinate, supervisor observation of the subordinate's job performance may be limited. Also, due to rapidly changing technology, the superior may not have enough understanding of the subordinate's actual on-the-job responsibilities to adequately rate his performance. Increasing automation in the workplace tends to widen the 'understanding gap' for the superior who does not strive to stay current in today's dynamic business world.

2. Peer Evaluators

Peer evaluators are those individuals who work at the same organizational level as the person rated. Many organizations avoid using peer evaluations, dismissing the technique as a 'popularity contest'. Peer evaluator-evaluatee friendship is seen as biasing the validity of this technique. This may be due to the perception that friends tend to minimize or overlook one another's shortcomings and only elevate good points, or mistake pleasing personal attributes for indicators of high job performance. Recent studies (e.g. Klimoski and London [Ref. 17], and Love [Ref. 18]) have shown that evaluation validity is not significantly affected by friendship bias, and that in some circumstances, peer evaluation appears to offer great benefits to an evaluation program.

3. Disinterested Party Evaluators

Disinterested parties can possibly be obtained within the organization or outside. They may come from any organizational level so long as they have no vested interest in the outcome of their evaluations. Some organizations bring in outsiders to perform this function, feeling that lack of personal contacts within the organization will allow a more objective evaluation.

A problem which may occur with disinterested party evaluators is that, aside from having no vested interest in the evaluation outcome, they may also have limited insight into the factors which indicate good job performance. As noted in supervisor evaluation, the evaluator who does not stay current on the the technology of the workplace is not likely to deliver as good a performance evaluation as one who is more familiar with that technology. In addition, outsiders brought in to perform evaluations may not fully grasp factors such as organizational politics and interpersonal relationships which can greatly influence overall job performance.

E. DISCUSSION

Each evaluation source has unique characteristics, as well as similarities with each of the other sources, in providing evaluation information. Though introduction of evaluator errors is fairly comparable for superior and peer evaluations [Ref. 19], studies have shown that rating sources differ in their perceptions of performance [Ref. 17]. This difference in perceptions is related to dimensionality.

Dimensionality is the quality of an evaluation area possessing different elements or dimensions. For instance, if one examined the broad area of secretarial job performance, many individual dimensions could be identified, such as typing speed, typing accuracy, shorthand ability, organization, ability to speak effectively on the telephone and many others. These dimensions comprise the evaluation area called secretarial job performance.

Not all evaluation sources use the same set of dimensions in conducting evaluations. As an example, consider an evaluation of worker performance performed by a worker's superior and a peer. The superior, being very goal oriented, rates the worker's clerical performance according to how many pages are typed per hour assuming, perhaps incorrectly, that quantity of pages typed also indicates quality. The peer, who must correct any errors made by the worker, is concerned with quality of output. Different sources exhibit different perceptions of performance. Neither view is necessarily wrong, but this illustrates the differences that may exist between evaluation sources. Holzburg [Ref. 19] has found a consistent outcome of dimensional analysis of superior and peer evaluations is that evaluation sources determine the primary dimensionality of the evaluations. What this means to the evaluatee is that performance grades received may be due more to the evaluation source than the job performance.

The following sections discuss some of the error sources which may cause evaluators to commit errors and methods of reducing various errors to provide more accurate evaluations.

F. ERROR SOURCES

Many factors contribute to evaluator error. Though often grouped under the general heading of bias, specific factors have been investigated by a variety of study groups as a way of ensuring objective and valid evaluations. This section looks at several of the factors contributing to evaluator error, and the next section discusses some methods suggested for reducing these errors.

1. Social Interaction

Social interaction, or friendship bias, is often cited as a reason for avoiding peer evaluations. As previously noted, this bias is thought by many organizations to adversely affect peer evaluations. This bias is also seen in superior evaluations, but judging from the number of organizations which use superior evaluators as a primary means of evaluation, the effects may not be considered as severe. This is not to say that superior evaluation biases are actually less severe than those biases found in other evaluation sources. The biases may be just as bad, but the superior's position tends to lend a degree of credibility to his or her judgements, deserved or not.

2. Evaluator Inexperience

Evaluator inexperience and lack of training in evaluation procedures tend to contribute to halo and leniency errors [Ref. 20]. Poorly defined measures force the inexperienced evaluator to make interpretations which, due to limited background, may not accurately reflect performance. Closely associated with this idea is the evaluator's effectiveness on the job. Low evaluator effectiveness correlates strongly with low evaluation accuracy.

3. Role Conflict

A strong factor contributing to evaluator error is the role conflict experienced by many evaluators. Dayal has noted:

> The manager has to accept the responsibility to judge the performance of other people. Often this responsibility is hesitantly taken because he feels uncomfortable in his role as judge. [Ref. 21:p.29]

One effect of this evaluator discomfort is that evaluation results tend to group near the upper end of the rating scale [Ref. 21]. A possible reason for this effect is that giving low ratings may result in slower promotion or even firing of an employee, for which the evaluator giving the ratings may feel responsible. Ratings at the high end of the scale reduce the probability that employees will experience layoffs or slower promotion and the evaluator will feel less responsible if such actions do occur.

4. Evaluator Knowledge of Evaluation Purpose

As previously stated, Scriven [Ref. 11] has suggested that evaluator knowledge of the evaluation purpose may be another nonperformance factor influencing the actual performance rating received. A study by Gallagher [Ref. 22] investigated whether ratings of performance varied when evaluators were given different purposes for the evaluations. The results support Scriven's contention. Gallagher's discussion of the results concludes "...that a single performance evaluation should not be used for different purposes since the stated purpose of the evaluation can affect the actual performance rating." [Ref. 22:p.38]

G. ERROR REDUCTION TECHNIQUES

Many techniques are available to help reduce evaluator error. These techniques have been investigated by various evaluation researchers (e.g. Bernardin [Ref. 23], Wiley and Jenkins [Ref. 24], and Scott [Ref. 20]) and some suggested solutions are presented here.

1. Evaluator Training

Bernardin, in a study of comprehensive vs. abbreviated evaluator training programs found that evaluators "...trained on error prior to observation and who used the scales to maintain observational diaries had significantly less leniency error and halo effect than all other groups." [Ref. 23:p.302] In this study comprehensive training was a one hour session consisting of definitions, graphic illustrations and examples of halo error, leniency and central tendency was presented to students who were acting as evaluators of peer performance. The trainees were also given data to evaluate in terms of the errors, and the evaluations were discussed. Abbreviated training was a five minute session with definitions of the error types and a single illustration of each.

The results of this study indicated that the psychometric quality for those who underwent comprehensive training was superior to those who received abbreviated training at the first rating period, and both training groups were superior to the control (untrained group). Another result was that the positive effects of the training programs were virtually nonexistent after one additional rating period. [Ref. 23] One might argue that for an organization contemplating a training program for supervisory personnel the above information may indicate that a comprehensive training program would lead to fewer evaluator errors than an abbreviated training program. As the effects of both training programs tends to rapidly diminish with time, however, a shorter training program regularly administered may deliver more positive effects in the long run.

2. Dimensional Analysis

As discussed previously, different evaluation sources perceive performance in different ways. To account for this, subjective evaluation areas should be examined by dimensional analysis. This analysis is used to investigate the many dimensions which comprise an evaluation area and considers the different combinations of dimensions used by various evaluation sources. Since each evaluation source tends to use different dimensions in performing evaluations [Ref. 25:p.473], dimensional analysis can provide insight into the particular concerns of the various sources. Klimoski and London [Ref. 17] present the example that supervisors may be less able to discriminate between items related to competence from those related to effort, whereas nurses rating themselves and peers can make that distinction. This would suggest that supervisors are more likely to consider effort as an indicator of competence than peers. By accounting for the dimensions used by various evaluation sources dimensional analysis can allow performance measures to be tailored according to the anticipated evaluation source, or it may be used after the fact to help explain ratings received in particular areas in light of the evaluation source.

3. Testing Evaluators

Wiley and Jenkins [Ref. 24] had 109 Air Force navigator students estimate qualifications needed to perform various Air Force tasks using an experimentally standardized task list and sets of five rating scales. Their estimates were aggregated and a consensus or pooled estimate group was formed. These students, after one month, again estimated qualifications and the students were scored by correlating their estimates with the key of pooled estimates. The study shows that evaluators who tend to agree with the consensus also tend to retest self-agreement. These evaluators also tend toward consensus agreement on later evaluations. [Ref. 24]

The above findings tend to suggest that a standardized test could be developed to rate potential evaluators. A consensus key which corresponds to the organization's view of performance would make it possible to select evaluators with corresponding views. This would help ensure organizational goals are being pursued by the evaluation process.

4. Reducing Subjectivity of Evaluation Measures

Performance appraisal systems are commonly regarded as being too subjective in nature, relying primarily on human judgement for gathering information pertaining to measures [Ref. 20]. Elimination of all factors which can not be objectively measured would naturally lead to minimal subjectivity. While this elimination may or may not be possible, it is possible to develop a system where the evaluator reacts to stimuli which are relatively free of subjective or irrelevant influences rather than stimuli which require the evaluator's judgement [Ref. 16:p.89-92]. The stimuli take the form of actual on-the-job incidents which the evaluator simply observes without interpretation. These incidents, or 'critical behaviors', represent actions normally associated with outstandingly successful or outstandingly unsuccessful task performance. The evaluator in this role acts as a reporter of actions rather than a judge who values actions [Ref. 20].

One problem associated with this method is the choice of critical incidents or behaviors. Some person or group of people must be designated to decide what incidents are to be used in evaluation. Providing a list of such incidents reduces the evaluator's need to exercise personal judgement in conducting evaluations.

H. SUMMARY

This chapter has investigated the evaluator as part of the scheme of evaluation. The concepts of objectivity and validity were introduced and explained as they pertain to evaluation. Sources of evaluator error were then discussed. Evaluator errors were divided into variable and constant errors, and each of these areas was broken into specific error types. Various evaluator sources- superior, peer and disinterested party- were discussed with advantages and disadvantages of each source considered. A discussion of error sources, along with techniques to reduce these errors closes the chapter. The last section suggests that training and testing evaluators and taking measures to reduce the subjectivity of evaluation measures can have a significant effect in reduction of evaluator error.

The next chapter uses the information presented in Chapters II and III to analyze the MCCRES and offer some suggestions for identifying and controlling or controlling for potential evaluator bias.

IV. MCCRES

> The purpose of the Marine Corps Combat Readiness Evaluation System (MCCRES) is to provide a timely and accurate evaluation of the readiness of Fleet Marine Forces, including Reserve units, to accomplish assigned missions. [Ref. 26:p.I-A-1]

To achieve the objective of timely and accurate readiness evaluation, the MCCRES has been designed to allow observation of Marine units in simulated combat situations. It promotes use of a standardized evaluation process and reporting system to provide feedback to the evaluated unit indicating strengths and weaknesses in a combat readiness posture. This chapter focuses on the evaluation process in an attempt to identify areas where evaluators may commit errors or inject bias into the evaluation possibly leading to inaccurate readiness ratings. The general evaluation approach and structure of the MCCRES are discussed first, followed by an investigation of potential sources of error. The final section discusses some solutions to minimize the effects of evaluator bias.

A. APPROACH

The MCCRES approach to evaluation may be compared with the Professional Review (Accreditation) Approach forwarded by House [Ref. 1]. It is an evaluation system conceived within the Marine Corps, graded by Marines and using standards developed by Marines. As such, it closely parallels the Professional Review Approach. In this approach, a particular profession sets standards of performance for itself and conducts internal evaluations. The reasoning for the internal evaluations is that members of that profession are considered experts in that field.

In choosing evaluators to perform MCCRES evaluations, it is desireable that evaluators have recently served successfully in a billet relating to the function they are to observe. This means, for example, that a Rifle Company evaluator should have recently served successfully as a Rifle Company commander. Successful recent billet performance increases the probability that evaluators will recognize adequate mission performance.

B. STRUCTURE

The MCCRES evaluation structure is a four-tiered hierarchy as shown in Figure 4.1. Of particular importance to this discussion are the bottom two layers--the Tactical Exercise Controller (TEC) and the Evaluators. It is here that mission performance is observed, analyzed and reported.

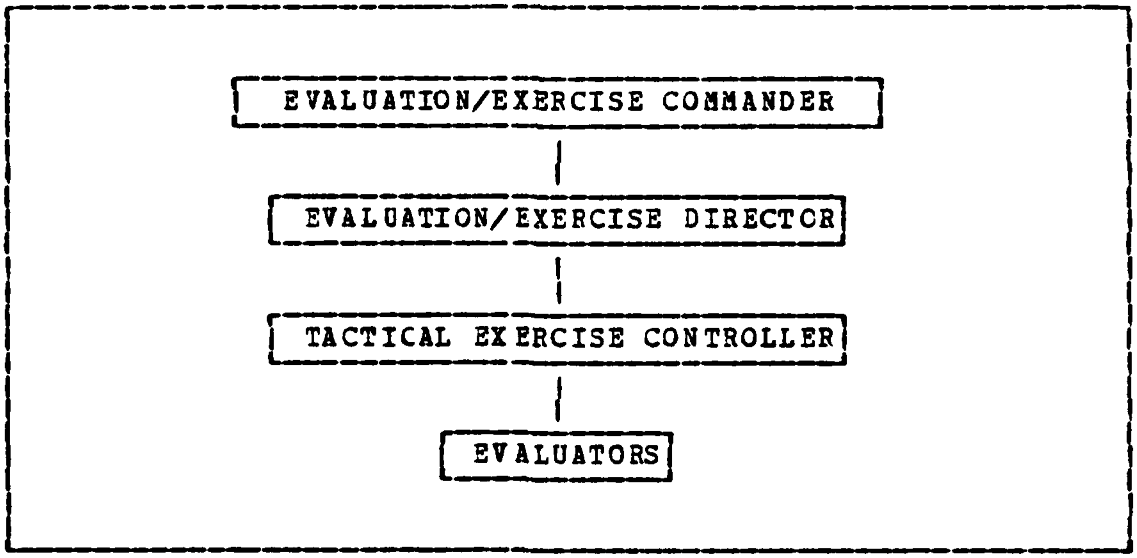

EVALUATION/EXERCISE COMMANDER
EVALUATION/EXERCISE DIRECTOR
TACTICAL EXERCISE CONTROLLER
EVALUATORS

Figure 4.1 MCCRES Evaluation Structure.

1. Tactical Exercise Controller (TEC)

The TEC compiles and analyzes the results of the evaluations which have been submitted via the evaluator's data sheets and submits a formal report to the Exercise Director. Among the TEC's duties and responsibilities are determination of specific Mission Performance Standards to be tested, extensive and detailed training of evaluators, development and control of intelligence play throughout the problem, and organization of the Tactical Exercise Control Group to plan and conduct the exercise. The TEC relies on the evalautors to report exercise progress and mission performance of the evaluated units. The former information is received primarily via radio communication while the latter arrives in the form of evaluator data sheets.

2. Evaluators

Evaluators have three main roles in the MCCRES:

1. Exercise controllers to ensure the exercise proceeds as planned.
2. Umpires to resolve disagreements between exercise and aggressor forces.
3. Performance evaluators to observe task performance as related to Mission Performance Standards being graded.

As an exercise controller, evaluators work as an extension of the will of the TEC. They may increase or decrease the operational tempo of the problem through the use of such items as aggressor forces, intelligence reports or simulated fires. They may create situations which require reaction by the evaluated unit by insertion of prescribed events into the play of the tactical problem. Action observed at this level is provided to the TEC primarily by radio to assist the TEC in determining if the exercise pace is satisfactory.

As umpires, evaluators are tasked with resolution of disagreements which may occur between evaluated units and aggressor forces. For example, if an evaluated unit was ambushed by an aggressor force, an evaluator would make a determination as to the outcome of the ambush and assess casualties accordingly.

In the role as performance evaluators, evaluators observe unit performance of prescribed tasks and make a determination as to the unit's ability to satisfactorily carry out the task. These determinations are recorded as "YES", "NO" or "NOT APPLICABLE" marks on the evaluator data sheet. A mark of "YES" denotes that all facets of a particular requirement were met. Conversely, a "NO" mark shows that all portions of a requirement were not met. "NOT APPLICABLE" areas are those not tested or which do not apply to the scenario at hand.

Having discussed the general roles of the evaluator, two topics are presented to help explain how MCCRES evaluators are organized and what measures are used in making a determination of combat readiness. The first, Senior Evaluators, explains the duties and relationships of this MCCRES member to the rest of the evaluators. The second, Mission Performance Standards, looks at the composition of the measures used in conducting the MCCRES.

a. Senior Evaluators

Each unit evaluated has a senior evaluator who conducts a post exercise wrap-up and compiles the data sheets from all subordinate evaluators. At this wrap-up, resolution of each "YES", "NO" and "NOT APPLICABLE" rating is made for each requirement tested. This resolution of the evaluator's data sheets results in "YES", "NO" or "NOT APPLICABLE" ratings for each requirement as it pertains to the entire unit. The senior evaluator provides his data sheets to the TEC for compilation and further use by the TEC. An assessment of "COMBAT READY" or "NOT COMBAT READY" for the entire unit is also also passed to the TEC by the senior evaluator.

The senior evaluator's relationship with other evaluators is a senior-subordinate type. Senior by position and generally by military rank, the senior evaluator is in charge of the evaluation team and is responsible for evaluating the performance of the entire unit being evaluated. The senior evaluator is appointed by name by the Exercise Director (an officer senior to the commander of the organization being evaluated) and as such, maintains an independent relationship to the organization being evaluated. Other members of the evaluation team, subordinate to the senior evaluator, are responsible for evaluating the subordinate units (both organic and attached) and other organizational functions (such as command and control and fire support coordination) of the overall unit being evaluated.

b. Mission Performance Standards

Mission Performance Standards (MPS's) are standards of task performance used in MCCRES. Each standard is composed of various tasks. For example, the MPS Continuing Actions By Marines is composed of twelve tasks such as Discipline, Dispersion, Security and Casualty Handling. These tasks are further divided into conditions and requirements. Conditions specify the circumstances under which requirements must be performed and provide recommendations to the evaluator concerning time and space limitations which may be imposed on the evaluated unit. Requirements are specific actions which must be performed or behaviors which must be demonstrated in the accomplishment of a given task. The task Discipline, for instance, contains nine requirements ranging from Self Discipline and Weapons Maintenance

Discipline tc Hygenic Discipline. Requirements which may need further information to guide evaluators in the determinaticn cf satisfactcry performance are provided with Key Indicators (KI's) of performance. KI's are an attempt to provide an cbjective foundaticn upon which to base an evaluator's judgement of satisfactory requirement performance. They should provide specific, measureable actions or behaviors which must be present for the requirement tc be successfully completed.

Consider the KI for the requirement Weapons Maintenance Discipline. "Marines take care to clean their weapcns, both individual and crew served, daily. Weapons are safeguarded. Care of weapons enforced by leaders." The KI tells what is to be dcne (clean weapons, both individual and crew served), when it is to be done (daily), who does it (Marines), and who supervises (leaders). KI's for cther requirements provide similar types cf information to make requirements more objectively measureable by the evaluator.

C. POTENTIAL PROBLEMS

This section discusses the areas in which evaluators may inject bias into the MCCRES. The discussion is presented in three parts: Senior evaluatcr influence, other evaluator bias and MPS problems. Some general solutions to these problems are suggested here with more specific solutions presented in the follcwing section.

1. Senior Evaluator Influence

The senior evaluatcr can inject bias in two major ways. First, as the senicr member of the evaluation team, he or she sets the tcne for the other evaluators. If the senior evaluator prcjects a hard-line, "by the book" approach toward the evaluation, evaluators may tend to view task requirements with little flexibility. On the cther hand, in a situation where the senior evaluator projects a less rigorous attitude toward the evaluation, evaluators may tend to view task requirements less rigidly. As a result of evaluator perceptions of the senior evaluator's wishes, the evaluation delivered may be biased.

The second major way in which the senior evaluator may inject bias is in the resolution of other evaluator's ratings. These ratings are obtained from the data sheets of the other evaluators. The senior evaluator depends upon the observations made by the other evaluators to provide data which accurately reflects the performance of the entire unit. Depending on the senior evaluator's perceptions of the other evaluators' competence and on his own perception of successful task completion, the senior evaluator's data for the TEC may or may not accurately reflect the overall unit's abilities. As an example, suppose an infantry battalion conducted an attack on an aggressor force and that two of the companies performed extremely well while one company performed poorly. If, in the senior evaluator's opinion, the offending company's performance was not critical to the entire unit's mission performance, a rating of "YES" could be delivered for the battalion for the task "ATTACK" as it pertains to the entire unit. [Ref. 26:p.I-C-8] On the other hand, if the senior evaluator felt the one company's performance was such that it negated the accomplishments of the other two companies, a rating of "NO" could conceivably be returned for the battalion for the task "ATTACK" as it pertains to the entire unit. The senior evaluator made a decision based on personal judgement, possibly reflecting the unit's mission performance inaccurately.

2. Other Evaluator Biases

The evaluators who observe task performance and report to the senior evaluator are presented with a continuing opportunity to inject bias into the MCCRES. The discussion of the areas where these evaluators may inject bias is organized in two groups: errors and evaluator sources.

a. Errors

Evaluator bias manifests itself as any deviation from the objective 'truth' concerning an evaluated unit's performance. In this respect, bias may be regarded as an error of leniency, strictness or halo effect. The first two errors result in ratings which are respectively too "easy" or too "hard", while the last error tends to cause ratings to group around one value on the rating scale. To illustrate, consider an evaluator rating the requirement Equipment Maintenance. The first portion of the KI for this requirement states "Vehicles, generators, etc., are given close attention by the Marines assigned to operate them." *[Ref. 26:p.II-A-6]* The lenient evaluator may consider visual observation each four hours constitutes close attention, while a strict evaluator considers maintenance conducted every other hour as an indicator of close attention. If a Marine is observed by these two evaluators checking his assigned equipment at strict four hour intervals because that is what the operating manual calls for, he will receive a different rating from each of the evaluators. In this case, the second evaluator has injected bias by committing the error of strictness.

As an illustration of halo error, suppose an evaluator is rating a unit on a task which contains five requirements. At the outset of the observation period, the unit was particularly outstanding in carrying out the first requirement. Based upon the outstanding performance the evaluator expects similar performance for the other requirements of the task. Such expectations may influence the evaluator to "see" only outstanding performance. Mistakes and poor performance are viewed with the attitude that "...they really know better, they just weren't paying attention today...". As a result of this attitude, a "YES" rating is delivered for the entire task, even though not all requirements were successfully completed. This evaluator has committed a halo error since the rating has been influenced by the outstanding performance of only one requirement of the entire task. It must be noted that this error can also be observed in the opposite sense, that is a particularly bad observation can bias the evaluator to view an entire task unfavorably.

b. Evaluation Sources

In the previous discussion of the three main sources of evaluation--superior, peer and disinterested party--it was shown that the first two sources demonstrate fairly comparable error introduction but may vary greatly in perceptions of task performance. This difference in perception is related to the dimensionality of the task being evaluated. In the context of MCCRES this means that superiors may not perceive task performance in the same way as peers. The last evaluation source, the disinterested party, brings with it the potential problem of not understanding the process being graded.

Many of the potential problems associated with various evaluation sources are diminished by two MCCRES stipulations concerning evaluators. The first stipulation is that evaluators should have recently served a successful tour in a billet related to the one they are evaluating. A key word in this stipulation is _recently_. Since billets in the Marine Corps have ranks associated with them, the differential dimensionality of senior and peer evaluators is limited by ensuring evaluators have _recently_ filled a billet similar tc the one they are evaluating. In other words, an evaluator who has recently served in a billet similar tc the one he is evaluating is more likely to recognize those task dimensions which indicate successful task performance than an evaluator who has not recently held such a position.

Besides the problems associated with differential dimensionality between evaluation sources, social interaction between sources and the evaluated unit can be problematic. Both seniors and peers within an organization tend to interact in formal as well as informal ways. This informal or social interaction may be carried into the evaluation as a bias. The second stipulation states "...it is desireable that evaluators be obtained from adjacent commands not directly related to the organization being evaluated." [Ref. 26:p.I-C-9] This may result in a reduction cf bias created by social interaction. This reduction is due to decreased daily interaction between members of adjacent units as compared to daily interactions among members cf a single unit.

3. Mission Performance Standards

All of the evaluation sources have one thing in common: they use the Mission Performance Standards to evaluate unit combat readiness. A potential problem associated with the MPS's is their subjectivity. This subjectivity permits evaluator interpretation of standards which may result in biased evaluations.

To determine the extent of the MPS's subjectivity, the requirements for the MPS's Continuing Actions By Marines, Command And Control and Fire Support Coordination were examined. The criterion used to determine the subjectivity of a requirement was the ability of the requirement to be quantified. If the requirement was expressed in terms which are physically measureable, such as units of time or distance, then it was considered objective. Requirements containing phrases which require interpretation by the evaluator, such as "...close attention...", were considered subjective. The meaning of these requirements can depend upon theevaluator's interpretation of the requirement's wording.

Of the 243 requirements for the above MPS's, 15 were found to be susceptible to evaluator interpretation. This is approximately 6.2 percent of the requirements for these three MPS's. These 15 requirements contain phrases such as "...close attention..." or "...processed with speed..." to describe satisfactory requirement performance. Without clear guidance as to what constitutes "close attention" or processing "with speed", different evaluators may interpret the requirement to have different meanings. This difference in interpretation means that two evaluators observing a particular requirement being performed could return different ratings of requirement performance, depending on how the requirement is interpreted. For each of the 15 requirements, the requirement number and the subjective phrase contained in the requirement is listed in Table II.

D. POTENTIAL PROBLEMS PERCEIVED BY FIELD USERS

Six Marine officers attending the Naval Postgraduate School were interviewed to gain an insight into potential MCCRES problems as perceived by users in the field. The six officers ranged in grade from O-2 to O-4 and represented MOS's 0302 (Infantry Officer) 1302 (Engineer Officer) 7562 (Pilot HMM CH-46) and 7587 (Airborne Radar Intercept Officer, F4N/J/S). The interview consisted of three questions:

1. Do you feel that an evaluator can affect a MCCRES evaluation through personal bias?
2. How is this bias input?

3. In what areas do you feel bias is most likely to occur?

The results of these interviews demonstrated that there was close agreement on each of the questions across both MOS and grade. All interviewees felt that an evaluator could affect a MCCRES evaluation through personal bias. This bias was seen as being input through evaluator interpretation of performance criteria. These criteria take the form of task requirements. Responses to the last question indicate field users felt bias is most likely to occur in those areas to which numerical measures are not easily attached. They felt areas which lend themselves to quantifiable measurement are less likely to contain evaluator bias than non-quantifiable areas.

TABLE II

MPS Requirements Susceptible to Evaluator Bias

Requirement Number	Subjective Phrase
2A.1.1.3	"close attention"
2A.1.1.4	"orderly and organized fashion"
2A.1.1.7	"exhibit restraint"
2A.1.1.8	"light use to a minimum"
2A.1.8.6	"COMSEC material safe-guarded"
2A.1.11.14	"processed with speed"
2A.2.7.2	"provided with security"
2A.2.8.2	"safeguards classified material"
2A.2.9.5	"neat and orderly"
2A.2.9.6	"dispersed to reduce vulnerability"
2A.2.10.5	"dispersed"
2A.3.4.5	"closely monitors"
2A.3.4.7	"timely manner"
2A.3.5.3	"accurate plots"
2A.3.5.7	"closely monitors"

Comparison of potential problems with MCCRES as perceived by the sample of field users to the potential problems outlined in the previous section shows that the field users' perceptions are a subset of the potential problems discovered through analysis of the MCCRES.

E. RECOMMENDED SOLUTIONS

The problems discussed in the previous two sections demonstrate the variety of ways in which an evaluator may introduce bias into a MCCRES. In order to minimize bias input, three possible solutions to the bias problem are forwarded. These solutions are evaluator training, evaluator testing and quantification of subjective MPS requirements.

1. Evaluator Training

As previously noted, evaluator training has proved to be an effective tool in reduction of evaluator error. Bernardin [Ref. 23] showed that evaluators receiving comprehensive training show greater error reduction results than evaluators receiving limited training. Both of these groups show less error than evaluators who have received no training.

Current MCCRES standards task the TEC with conducting extensive and detailed training of evaluators. In the experience of several officers attending the Naval Postgraduate School, who were questioned concerning evaluator training, this training is geared toward educating the evaluator on the exercise scenario with no specific mention of the errors which evaluators typically commit. By making MCCRES evaluators aware of the errors typically committed by evaluators, the MCCRES evaluators are less likely to commit these errors, reducing biased input. An evaluator training package addressing both scenario development and possible evaluator error should be created to more fully exploit the potential of comprehensive evaluator training outlined by Bernardin [Ref. 23].

Another aspect of evaluator training is ensuring potential evaluators are well-versed in the areas they are chosen to evaluate. Choosing knowledgeable evaluators tends to increase the probability that those factors which indicate successful task performance are considered during the evaluation.

One method to ensure trained, knowledgeable evaluators for MCCRES evaluations is formation of a formal MCCRES evaluation team. By choosing team members who have demonstrated proficiency in their MOS's and keeping them current in both their MOS's and evaluation techniques through training, a skilled cadre of evaluators can be assembled.

Some of the advantages of forming a formal MCCRES evaluation team are minimization of evaluator training costs, minimization of social interaction with evaluated units and a more standardized evaluation base. Evaluator training costs are minimized since the same evaluators are frequently used. Though training affects diminish rapidly with time, retraining for each successive evaluation could demonstrate a learning curve, reducing costs over time. Social interaction is minimized due to lower daily contact with evaluators, as opposed to the interaction which occurs among adjacent commands. The last factor, standardization of the evaluation base, results from the continuity of the formal evaluation team.

A MCCRES evaluation team could be composed of personnel from units such as Division Schools, or it could reside outside the active duty forces at a Reserve unit, since the MCCRES is to evaluate both active and reserve forces. Having reserves evaluate MCCRES would also offer the additional benefit of keeping the reserve up to date and strengthening the tie between active and reserve forces in the Marine Corps.

2. Evaluator Testing

Evaluator testing can be seen as a method of both controlling and controlling for evaluator bias. In the former case, a test can be constructed which would indicate the areas in which a prospective evaluator demonstrates bias. By testing a number of these prospective evaluators, those who demonstrate little or no bias could be chosen to conduct MCCRES evaluations, thereby minimizing the likelihood of evaluator bias input. For instance, consider a test in which evaluators are graded according to their agreement with an answer key. Further, suppose the answer key is composed of the pooled answers of a group of "unbiased" evaluators. As suggested by Wiley and Jenkins [Ref. 24:p.217], evaluator agreement with the key can be used to predict the likelihood of evaluator bias. Those evaluators showing close agreement with the key of "unbiased" answers can be chosen to perform evaluations.

The same test, analyzed differently, can be used to control for evaluator bias. For instance, the results of the test are analyzed to discover in which areas an evaluator's biases exist. From this analysis a "bias profile" could be constructed which could allow evaluation results to be "standardized". For example, assume a MCCRES evaluator's bias profile showed significant deviation toward strictness in the area of discipline. During the conduct of a MCCRES evaluation a senior evaluator notes this evaluator's data sheet has a "NO" rating for many of the requirements of the task DISCIPLINE. The senior evaluator, knowing that this evaluator tends to be particularly strict in evaluating discipline, may wish to obtain additional performance information concerning the unit evaluated, since the evaluator's ratings may not accurately reflect the unit's actual performance.

3. Quantification of MPS's

The last method of controlling evaluator bias is quantification of subjective MPS requirements. This quantification, as Scott [Ref. 20] suggests, reduces the evaluator's task from interpreting MPS requirements and comparing task performance with this interpretation to reporting whether task performance meets the requirements. For example, instead of trying to decide how fast the phrase "...process with speed..." is, reporting whether the unit was able to "...process within two hours..." is less open to interpretation. The more concrete the requirement, the less evaluator interpretation that will take place in grading, resulting in reduced evaluator bias. Some of the quantifications may be less concrete than others. Some requirements may be constructed in terms of ranges of acceptable performance for differing tactical scenarios. Still, the ranges serve to bound the amount of interpretation required by the evaluator.

F. CONCLUSIONS

In the introduction of this paper two questions are posed. The first asks if factors of the MCCRES evaluation which are subject to evaluator bias can be identified, and the second asks how these factors can be controlled or controlled for. It has been shown that areas in which evaluators may bias the MCCRES can be identified and comprise three basic areas: senior evaluator influence, other evaluator bias and MPS interpretation.

As for methods of controlling or controlling for these factors, three techniques were forwarded: evaluator training, evaluator testing and quantification of subjective MPS requirements. Each of these techniques has potential for controlling bias.

G. RECOMMENDATIONS FOR FUTURE RESEARCH

Discussion of the proposed solutions to the problem of evaluator bias did not address the cost to implement the solutions. A study of benefits and costs for each of the solutions would provide additional information as to the feasibility of the solutions. In addition, a detailed study of the proposed solutions would be likely to point out several methods of implementation for each, possibly revealing still other solutions not addressed in this thesis.

LIST OF REFERENCES

1. House, E. R., *Evaluating With Validity*, Sage Publications, 1980.

2. Levitan, S. A. and G. Wurzburg, *Evaluating Federal Social Programs*, W. E. Upjohn Institute for Employment Research, 1979.

3. Reiken, H. W., "Action for What? A Critique of Evaluative Research," in *Evaluating Action Programs: Readings in Social Action and Education*, ed. Carol H. Weiss, Allyn and Bacon, 1972.

4. Stufflebeam, D. L., W. L. Foley, W. J. Gephart, E. G. Guba, R. L. Hammond, H. O. Merriman, and M. M. Provus, *Educational Evaluation and Decision Making*, F. E. Peacock Publishers, Inc., 1971.

5. Anderson, S. B., and S. Ball, *The Profession and Practice of Program Evaluation*, Jossey-Bass Inc. Publishers, 1978.

6. Tracey, W. R., *Evaluating Training and Development Systems*, American Management Association, 1968.

7. Langston, J. H., "OEO Neighborhood Health Centers: Evaluation Case Study", in *Social Experiments and Social Program Evaluation*, eds. J. G. Abert and M. Kamrass, p.107-121, Ballinger, 1974.

8. Tyler, R. W., *Basic Principles of Curriculum Instruction*, University of Chicago Press, 1950.

9. Drucker, P. F., *The Practice of Management*, Harper and Brothers, 1954.

10. Shuster, F. E., and A. F. Kindall, "Management by Objectives: Where We Stand- A Survey of the Fortune 500," *Human Resource Management*, v.13, no. 1,Spring 1974.

11. Scriven, M., "Goal Free Evaluation," in *School Evaluation*, ed. E. R. House, McCutchan, 1973.

12. Eisner, E., *The Educational Imagination*, McMillan, 1979.

13. Wolf, R. L., "The Use of Judicial Evaluation Methods in the Formulation of Educational Policy," *Educational Evaluation and Policy Analysis 1*, May-June 1974.

14. Barrett, R. S., *Performance Rating*, Science Research Associates, 1966.

15. Scriven, M., "Objectivity and Subjectivity in Educational Research," in *Philosophical Redirection of Educational Research*, ed. L. G. Thomas, National Society for the Study of Education, 1972.

16. Cummings, L. L., and D. P. Schwab, *Performance In Organizations*, Scott, Foresman and Company, 1973.

17. Klimoski, R. J. and M. London, "Role of the Rater in Performance Appraisal", *Journal of Applied Psychology*, vol. 59, no. 4, p.445-451, 1974.

18. Love, K. G., "Comparison of Peer Assessment Methods: Reliability, Validity, Friendship Bias, and User Reaction," *Journal of Applied Psychology*, vol. 66, no. 4, p. 451-457, 1981.

19. Holzbach, R. L., "Rater Bias in Performance Ratings: Superior, Self-, and Peer Ratings," *Journal of Applied Psychology*, vol. 63, no. 5, p. 579-588, 1978.

20. Scott, R. D., "Taking Subjectivity Out of Performance Appraisal," *Personnel*, p. 45-49, July-August 1973.

21. Dayal, I., "Some Issues in Performance Appraisal," *Personnel Administration*, p.29-35, January-February 1969.

22. Gallagher, M. C., "More Bias in Performance Evaluation?", *Personnel*, p. 35-40, July-August 1978.

23. Bernardin, H. J., "Effects of Rater Training on Leniency and Halo Errors in Student Ratings of Instructors," *Journal of Applied Psychology*, vol. 63, no. 3, p. 301-308, 1978.

24. Wiley, L., and W. Jenkins, "Selecting Competent Raters," *Journal of Applied Psychology*, vol. 48, no. 4, p. 215-217, 1964.

25. Guion, R. M., *Personnel Testing*, McGraw-Hill, 1965.

26. *Marine Corps Order 3501.2*, *Vol.II*, 9 December, 1977.

INITIAL DISTRIBUTION LIST

No. Copies

1. Defense Technical Information Center — 2
Cameron Station
Alexandria, Virginia 22314

2. Library, Code 0142 — 2
Naval Postgraduate School
Monterey, California 93940

3. Assistant Professor Kenneth J. Euske, Code 54EE — 2
Department of Administrative Sciences
Naval Postgraduate School
Monterey, California 93940

4. Lieutenant Colonel Joseph F. Mullane, USMC — 5
Code 0309
Marine Corps Representative
Naval Postgraduate School
Monterey, California 93940

5. Department Chairman, Code 54 — 1
Department of Administrative Sciences
Naval Postgraduate School
Monterey, California 93940

6. Computer Technology Programs — 1
Code 37
Naval Postgraduate School
Monterey, California 93940

7. Commandant of the Marine Corps (Code POR) — 1
Headquarters Marine Corps
Washington, D.C. 20380

8. Captain George M. Wheeler, USMC — 2
705 Fifth Street
Marietta, Ohio 45750